

ISSN: 1286-4892

■ Ababacar MBENGUE 2010

Faut-il brûler les tests de signification statistique ?

M@n@gement, 13(2), 99 -127.

accepté par Louis Hébert

M@n@gement vol. 13 no. 2, 2010, 99-127

Faut-il brûler les tests de signification statistique ?

Ababacar MBENGUE

Reims Management School
ababacar.mbengue@univ-reims.fr

La démarche d'inférence occupe une place centrale dans la recherche en management. Très souvent, le chercheur est amené à tirer des conclusions ou à procéder à des généralisations à partir de ses observations ou de ses résultats. Dans certains cas, la statistique peut lui permettre de le faire de manière rigoureuse à travers la statistique inférentielle qui est la démarche d'inférence par laquelle le statisticien teste la généralisation d'une information collectée sur un échantillon à l'ensemble de la population dont est issu cet échantillon. Les tests statistiques sont ainsi au cœur de la statistique inférentielle et, par suite, de la démarche d'inférence. Pourtant, depuis leur introduction, les tests de signification statistique font l'objet de critiques vives et récurrentes portant aussi bien sur leur nature que sur leur rôle (Nickerson, 2000). La critique des tests de signification statistique est présente depuis longtemps dans pratiquement tous les champs disciplinaires... à l'exception notable du management qui commence tout juste à aborder la question (Mbengue, 2007 ; Schwab et Starbuck, 2009). Dès lors, l'objectif principal de cet article est précisément de mieux informer la communauté des chercheurs en management de l'existence de ces critiques des tests de signification statistique, d'en détailler le contenu et les enjeux et, surtout, de proposer des recommandations permettant d'améliorer le test d'hypothèses et, au-delà, la démarche d'inférence – statistique – dans la recherche en management.

Mots clés : inférence, méthodologie, statistique, signification statistique, tests statistiques.

Inference plays a central role in management research as researchers are frequently led to draw conclusions or make generalizations from their observations or results. In many cases, they are able to do this rigorously through inferential statistics, which is the process of inference whereby the statistician tests the generalization of information collected in a sample to the entire population the sample is from. Statistical tests are thus at the heart of inferential statistics and, consequently, the process of inference. However, since they were first developed, statistical significance tests have been the object of sharp and repeated criticism regarding both their nature and their role (Nickerson, 2000). Such criticism has been longstanding in virtually all disciplines, with the notable exception of management that is just beginning to address the issue (Mbengue, 2007, Schwab & Starbuck, 2009). The main purpose of this paper is to provide researchers in management with clear information about the controversy surrounding statistical significance tests, to detail the content and issues and, most importantly, to offer recommendations for improving the testing of hypotheses and beyond, in other words, the process of statistical inference in management research.

Keywords: *methodology, inferential statistics, statistical tests, statistical significance*

INTRODUCTION

Dès leur introduction, les tests de signification statistique (TSS) ont fait l'objet de multiples critiques portant à la fois sur leur nature et sur leur rôle. Et si quelques auteurs ont fourni une défense parfois intéressante de ces tests (Hagen, 1997 ; Mulaik, Raju et Harshman, 1997 ; Wainer, 1999), de nombreux autres ont appelé à leur abolition pure et simple (Hunter, 1997 ; Gill, 1999 ; Armstrong, 2007a, 2007b). De fait, peu de méthodes statistiques ont été aussi fortement et durablement critiquées que les TSS (Morgan, 2003). C'est ainsi que Rozeboom (1960 : 416) les a qualifiés de « sophisme », Bakan (1966 : 436) d'« exercice d'insouciance dans la conduite de recherche », Carver (1978 : 397) de « forme corrompue de la méthode scientifique », Krueger (2001 : 16) de « méthode erronée » et Schmidt et Hunter (2002 : 66) de « désastre ». Pourtant, en dépit de ces nombreuses critiques, les TSS continuent d'être aussi populaires qu'employés à mauvais escient (Finch, Cumming et Thomason, 2001 ; Morgan, 2003). De fait, les multiples recherches effectuées dans des champs disciplinaires aussi divers que la psychologie, la sociologie, le marketing, la comptabilité, les sciences de l'éducation, les sciences politiques, l'écologie, la prospective, la psychiatrie, etc. montrent de manière constante que les chercheurs méconnaissent les bases les plus élémentaires des TSS (Nelson, Rosenthal et Rosnow, 1986 ; Zuckerman, Hodgins, Zuckerman et Rosenthal, 1993 ; Mittag et Thompson, 2000). Alors, faut-il oui ou non brûler les TSS ?
Pour répondre à cette question, nous allons décrire brièvement la logique générale des TSS, présenter les principales critiques qui leur sont adressées puis proposer des recommandations permettant d'améliorer le test d'hypothèses et, au-delà, la démarche d'inférence – statistique – dans la recherche en management.

LOGIQUE GÉNÉRALE DES TESTS DE SIGNIFICATION STATISTIQUES

INFÉRENCE ET STATISTIQUE

La démarche d'inférence occupe une place centrale dans l'activité du chercheur qui est fréquemment amené à interpréter des résultats, à tirer des conclusions ou à procéder à des généralisations à partir de ses observations. Par exemple, dans quelle mesure les résultats obtenus lors d'une recherche sur un nombre limité de cas, d'individus, de groupes sociaux, d'organisations, des processus, etc. sont-ils également valables pour d'autres cas, individus, groupes sociaux, organisations, processus, etc. non inclus dans la recherche initiale ? Il s'agit là d'un trait essentiel de la démarche scientifique puisque l'intérêt d'une recherche est souvent largement tributaire du caractère plus ou moins généralisable de ses résultats même si, bien entendu, toute recherche ne vise pas des résultats de portée universelle. De fait, le chercheur se contente rarement de décrire des situations, des phénomènes ou des

M@n@gement vol. 13 no. 2, 2010, 99-127

données bruts. Très souvent, il est amené à partir d'éléments particuliers pour aboutir à des conclusions de portée plus générale. Autrement dit, il doit procéder à une inférence pour généraliser ses résultats.

On distingue généralement deux formes principales d'inférence : 1) l'inférence théorique ou généralisation analytique (Kennedy, 1979 ; Yin, 1984 ; Firestone, 1993 ; Smaling, 2003), dont l'objectif est de généraliser des propositions théoriques sur la base d'un raisonnement logique ; 2) l'inférence statistique qui est une forme de généralisation qui s'appuie sur les propriétés de la statistique mathématique. Bien que chacune de ces deux formes d'inférence occupe une place importante dans la recherche scientifique, c'est l'inférence statistique qui est au cœur des TSS en cela qu'elle permet au chercheur de tester la généralisation d'une information collectée sur un échantillon à l'ensemble de la population dont est issu cet échantillon. Cette inférence statistique mobilise deux classes d'outils : d'une part les méthodes de test qui seront examinées dans cet article et, d'autre part les méthodes d'estimation, ponctuelle et par intervalle, dont la présentation sort du cadre de notre propos.

En tout état de cause, toute généralisation statistique des résultats d'une recherche n'a de sens qu'en des termes probabilistes. Cela signifie que l'inférence statistique ne conduit jamais à des jugements certains mais, au contraire, uniquement à des jugements plus ou moins probables. Cela signifie aussi qu'elle ne peut pas éliminer complètement le risque d'erreur associé au choix de conférer à un phénomène une validité qui dépasse le contexte restreint dans lequel il a été observé. Pour autant, la démarche d'inférence statistique garde un intérêt tout à fait réel dans la mesure où elle peut contrôler ce risque d'erreur. Plus exactement, elle parvient à en faire une estimation théoriquement précise, indiquant au chercheur la probabilité qu'il a de se tromper en généralisant les conclusions de son étude. À cette fin, l'inférence statistique fait appel à la statistique descriptive et à la théorie des probabilités dont l'objet est l'étude des lois et des régularités qui régissent les phénomènes aléatoires. Cette théorie permet à l'inférence statistique de déterminer la probabilité qu'un phénomène observé sur un échantillon soit dû uniquement au hasard de l'échantillonnage, alors même qu'il serait inexistant dans la population tout entière.

L'inférence statistique s'effectue au moyen de tests d'hypothèses statistiques. Une hypothèse statistique est un énoncé quantitatif concernant les paramètres d'une population (Baillargeon et Rainville, 1978). On appelle paramètre d'une population un aspect quantitatif de cette population comme la moyenne, la variance, un pourcentage ou encore toute quantité particulière relative à cette population. Les paramètres d'une population sont généralement inconnus. Cependant, il est possible de les estimer de manière statistique à partir d'un échantillon issu de la population. Pour pouvoir procéder à une inférence statistique, le chercheur doit traduire son hypothèse de recherche en hypothèse statistique. Par exemple, une hypothèse de recherche peut être que, pour une population donnée d'entreprises, celles qui ont adopté des démarches de gestion de la qualité totale obtiennent des performances supérieures à celles qui ne l'ont pas fait. Pour passer d'une hypothèse de

recherche à son test au moyen de la statistique, il faut préalablement la traduire en hypothèse statistique. Les modalités de cette traduction de l'hypothèse de recherche en hypothèse statistique sont de ce fait cruciales pour la validité d'une inférence portant sur l'hypothèse de recherche. Pour autant, les théories des tests statistiques ne concernent que l'inférence portant sur l'hypothèse statistique et la question de la pertinence de la traduction de l'hypothèse de recherche en hypothèse statistique n'est pas de leur ressort (Poitevineau, 1998).
Une hypothèse statistique se présente traditionnellement sous la double forme d'une première hypothèse appelée hypothèse nulle et d'une seconde hypothèse appelée hypothèse alternative ou contraire. L'hypothèse nulle désigne en général les situations d'absence de changement ou d'écart par rapport à un statu quo ou encore d'absence de différence entre des paramètres. C'est de là que provient la dénomination d'hypothèse nulle (Kanji, 1993 ; Dodge, 1993). Très souvent, l'objectif du chercheur est de réfuter cette hypothèse nulle au profit de l'hypothèse alternative (Dodge, 1993 ; Sincich, 1996 ; Zikmund, 1994). L'hypothèse alternative est alors celle que le chercheur souhaite établir, celle à laquelle il croit (Sincich, 1996). L'hypothèse nulle et l'hypothèse alternative ou contraire sont incompatibles et décrivent deux états complémentaires de la nature. L'hypothèse nulle est généralement notée H0 et l'hypothèse alternative H1 ou Ha. On notera que les tests statistiques sont conçus pour la réfutation et non la confirmation d'hypothèses. En d'autres termes, ces tests n'ont ni l'ambition ni le pouvoir de prouver des hypothèses : ils permettent de montrer qu'une hypothèse ne peut pas être acceptée parce qu'elle est associée à un niveau de probabilité trop faible (Kanji, 1993).
Même s'ils sont généralement présentés comme une théorie unifiée, les TSS sont en fait une hybridation de deux paradigmes concurrents : celui du test de signification de Fisher et celui du test d'hypothèse de Neyman et Pearson (Hubbard et Bayarri, 2003). La distinction entre ces écoles de pensée concurrentes ne semble pas avoir occupé un grand rôle dans le débat sur l'utilisation des TSS, bien que certains auteurs expliquent une grande part des limites des TSS par cette hybridation originelle (Harlow, 1997 ; Gill, 1999 ; Hubbard et Bayarri, 2003).

TESTS ET ERREURS STATISTIQUES

L'évaluation de la validité d'une hypothèse statistique se fait au moyen d'un test statistique effectué sur des données issues d'un échantillon représentatif de la population étudiée. Ce test statistique est une procédure permettant d'aboutir, en fonction de certaines règles de décision, au rejet ou au non-rejet d'une hypothèse de départ, en l'occurrence l'hypothèse nulle. La forme des tests statistiques dépend du nombre de populations concernées (une, deux ou davantage). Dans un test statistique portant sur une seule population, on cherche à savoir si la valeur d'un paramètre θ de la population est identique à une valeur présumée. L'hypothèse nulle, qui est dans ce cas une supposition sur la valeur présumée de ce paramètre, se présente alors généralement sous la forme suivante :

$$H0 : \theta = \theta 0$$

où θ est le paramètre de la population à estimer et $\theta 0$ la valeur présumée de ce paramètre inconnu θ.
Quant à l'hypothèse alternative, elle pose l'existence d'une différence ou d'une inégalité. Par exemple, on peut faire l'hypothèse d'une supériorité de performance des entreprises qui planifient formellement. Dans un tel cas, le test statistique qui sera effectué est un test dit test unilatéral ou unidirectionnel à droite. Si l'hypothèse était celle d'une infériorité de performance des entreprises planificatrices, il faudrait effectuer un test unilatéral ou unidirectionnel à gauche. Enfin, si l'hypothèse formulée devenait simplement celle d'une différence de performance sans autre précision, il faudrait effectuer un test bilatéral ou bidirectionnel. Il apparaît ainsi que l'hypothèse alternative peut prendre trois formes différentes :

H1 : $\theta > \theta 0$ (unilatéral ou unidirectionnel à droite)
H1 : $\theta < \theta 0$ (unilatéral ou unidirectionnel à gauche)
H1 : $\theta \neq \theta 0$ (bilatéral ou bidirectionnel)

Les tests statistiques sont effectués dans le but de prendre une décision, en l'occurrence rejeter ou ne pas rejeter l'hypothèse nulle, H0. Mais parce que la décision est fondée sur une information partielle issue d'observations portant sur un échantillon de la population, elle comporte un risque d'erreur (Baillargeon et Rainville, 1978 ; Sincich, 1996 ; Zikmund, 1994). On distingue deux types d'erreurs dans les tests statistiques : l'erreur de première espèce, dont la probabilité est notée α, et l'erreur de seconde espèce dont la probabilité est notée β. Les observations de l'échantillon peuvent conduire à rejeter l'hypothèse nulle H0 alors que la population remplit effectivement les conditions de cette hypothèse. Le risque de première espèce, α, mesure cette probabilité de rejeter l'hypothèse nulle H0 alors qu'elle est vraie. Inversement, les observations de l'échantillon peuvent conduire à ne pas rejeter l'hypothèse nulle H0 alors que la population remplit les conditions de l'hypothèse alternative H1. Le risque de seconde espèce, β, mesure cette probabilité de ne pas rejeter l'hypothèse nulle H0 alors qu'elle est fausse. Une erreur de première espèce ne peut survenir que dans les cas où l'hypothèse nulle est rejetée. De même, une erreur de seconde espèce ne peut avoir lieu que dans les cas où l'hypothèse nulle n'est pas rejetée. Par conséquent, soit le chercheur ne commet pas d'erreur, soit il en commet, mais d'un seul type. Il ne peut pas commettre à la fois les deux types d'erreur. On appelle puissance d'un test statistique la probabilité $(1 - \beta)$ de rejeter l'hypothèse nulle H0 alors qu'elle est fausse. La puissance d'un test est d'autant plus grande que le risque de deuxième espèce β est faible. La plupart des logiciels d'analyse statistique fournissent la probabilité associée à la valeur observée de la statistique calculée ou seuil de signification observé plus connue sous le nom de valeur p (« p-value »). Il s'agit de la probabilité, calculée sous l'hypothèse nulle, d'obtenir un résultat aussi extrême que la valeur obtenue par le chercheur à partir de son échantillon (Dodge, 1993). L'hypothèse nulle H0 sera rejetée si la valeur p est inférieure au seuil de signification fixé α (Sincich, 1996).
Concrètement, les tests d'hypothèse permettent d'évaluer le risque d'erreur auquel le chercheur s'expose en conférant un caractère géné-

ral aux résultats de sa recherche. Par exemple, ayant constaté sur un échantillon de soixante entreprises cotées à Paris, une corrélation entre la taille et la mise en place de politiques de gestion des compétences, des chercheurs veulent déterminer si l'on peut raisonnablement conclure qu'un phénomène de même nature est également présent dans l'ensemble de la population des entreprises cotées à Paris. À cette fin, ils procèdent à un test statistique pour vérifier dans quelle mesure (avec quel risque d'erreur) la corrélation observée sur l'échantillon de soixante entreprises peut être généralisée à l'ensemble de la population des entreprises cotées à Paris. Précisément, cette démarche de test d'hypothèse permet de déterminer la probabilité qu'un résultat égal ou supérieur à celui qui a été obtenu sur l'échantillon soit dû uniquement au hasard de l'échantillonnage. Elle conduit donc à évaluer le risque d'erreur d'effectuer une généralisation abusive du résultat de l'échantillon pour la population.

LA CRITIQUE DES TESTS DE SIGNIFICATION STATISTIQUE

Fondamentalement, les TSS posent trois types de problèmes : 1) sur le plan statistique, ils présentent des faiblesses intrinsèques qui sont trop souvent ignorées ; 2) sur le plan technique, l'usage dominant de ces outils est généralement inadéquat ; 3) sur le plan philosophique, ils posent le problème plus général des limites de l'inférence et reposent in fine sur la conviction indémontrable que la connaissance du passé permet de prédire l'avenir (Cox, 1958 ; Krueger, 2001 ; Morgan, 2003).

DES PROBLÈMES D'ORDRE STATISTIQUE

Une hypothèse vraiment… nulle

Plusieurs éminents statisticiens ont reconnu de longue date que l'hypothèse nulle d'absence de différence n'est jamais vraie dans la population (Nunnally, 1960 ; Tukey, 1991). Parmi les nombreux exemples d'hypothèses nulles, on peut mentionner : une fréquence ou une proportion égale à zéro ; une corrélation égale à zéro ; un coefficient de régression égal à zéro ; une égalité entre deux ou plusieurs moyennes, etc. Selon Schwab et Starbuck (2009), de tels exemples sont très fréquents dans la recherche en management. De nombreux détracteurs des TSS mettent en cause l'utilité même d'une hypothèse nulle (H0) ponctuelle, c'est-à-dire une hypothèse attribuant au paramètre une valeur précise et non un intervalle. En effet, une telle hypothèse est pratiquement toujours fausse, ne serait-ce qu'à plusieurs décimales après la virgule (Oakes, 1986 ; Tukey, 1991 ; Morgan, 2003). Ainsi, toute différence, même infinitésimale, deviendra statistiquement significative pour autant que la taille de l'échantillon soit suffisamment grande. Par exemple, pour un test de différence de moyennes entre deux groupes, toute différence non rigoureusement nulle pourra être « rendue » significative pour autant que la taille des groupes soit suffisamment grande. Considérons l'exemple d'une recherche visant à tester, par une

expérimentation, l'hypothèse (en réalité vraie) selon laquelle une action de formation spécifique augmente la productivité des salariés. Le groupe expérimental (les salariés ayant suivi la formation) et le groupe de contrôle (les salariés n'ayant pas suivi la formation) représentent alors deux populations différentes. Le groupe expérimental présente une performance moyenne de 76 % et un écart-type de 4,3 et le groupe de contrôle une performance moyenne de 72 % avec un écart-type de 4,3. On peut démontrer qu'avec des échantillons représentatifs des deux populations, il faut une taille d'au moins 11 salariés par groupe pour obtenir une différence statistiquement significative entre les performances des deux groupes. Donc, si la recherche porte sur moins de 11 salariés par groupe, on obtiendra un résultat statistiquement non significatif alors qu'avec un nombre de salariés supérieur à 10 on aura un résultat statistiquement significatif. Dans cet exemple, le chercheur peut directement décider du caractère statistiquement significatif ou non significatif de ses résultats uniquement en contrôlant la taille de l'échantillon. Une autre façon de voir les choses est de partir des populations et non plus des échantillons. Il est très probable qu'à l'échelle des populations l'hypothèse nulle soit fausse (il y aura une différence, même infinitésimale). Dès lors, tout test conduira à un résultat significatif et l'information apportée par le test est donc quasi nulle. Cela conduit plusieurs détracteurs des tests à la conclusion de l'inutilité de recueillir des données et de procéder à de tels tests dont les résultats sont connus d'avance (Meehl, 1990 ; Krueger, 2001 ; Morgan, 2003).

Une démarche anti-scientifique ?

L'accent mis par les TSS sur l'hypothèse nulle (hasard, absence de différence, statu quo, absence d'effet, etc.) aurait conduit à affaiblir la démarche scientifique. En effet, celle-ci consiste essentiellement à confronter les données recueillies à l'hypothèse de recherche (H1), puis, quand elles semblent incompatibles avec H1, à envisager d'autres hypothèses (dont celle du hasard, éventuellement). Au contraire, dans le cas des TSS, l'hypothèse du hasard (H0) est généralement mise en avant ; elle est la première testée, quel que soit son (in)intérêt scientifique (Carver, 1978, 1993 ; Cohen, 1994). Par conséquent, l'hypothèse de recherche ne sera même pas examinée si le test est non significatif, alors qu'elle pourrait pourtant présenter une bonne compatibilité avec les données. Si le choix d'assimiler l'hypothèse nulle aux situations de hasard, d'absence de différence et d'effet, etc. présente de multiples avantages techniques comme la facilitation de certains calculs statistiques, il n'en demeure pas moins que cela éloigne de la préoccupation principale du chercheur qu'est sa question de recherche (souvent H1). Or, lorsqu'un TSS est utilisé, tout se passe comme si seuls deux modèles existaient ou méritaient d'être considérés : l'hypothèse nulle et l'hypothèse alternative (souvent celle du chercheur). Cela est illusoire dans la mesure où plusieurs autres modèles alternatifs sont possibles (Rozeboom, 1960 ; Lindsay, 1995 ; Morgan, 2003). Prenons l'exemple d'une hypothèse de recherche selon laquelle « la distribution de stock-options à des dirigeants accroît leur fidélisation ». L'hypothèse nulle s'exprime alors de la manière suivante : « la distribution de stock-options à des dirigeants ne modifie pas leur fidélisation ». Un test de

différence des moyennes entre deux échantillons de dirigeants ayant reçu et n'ayant pas reçu des stock-options révèle ensuite un taux de fidélisation nettement supérieur pour les dirigeants ayant reçu des stock-options qui se traduit par une différence de moyenne statistiquement significative. Le rejet de l'hypothèse nulle vaudra ici acceptation ou, du moins, « non-rejet » de l'hypothèse de recherche qui était celle de l'accroissement de la fidélisation due à l'attribution de stock-options. Or plusieurs phénomènes autres que l'attribution de stock-options (l'hypothèse de recherche) peuvent être à l'origine de la différence de moyennes observée, par exemple un biais méthodologique ou une erreur de mesure ou encore des variables comme le type d'entreprise, l'ancienneté dans l'entreprise ou le profil personnel des dirigeants. Comme le signale Carver (1978), en cas de rejet de l'hypothèse nulle, il faudrait pouvoir écarter toutes les explications rivales à l'hypothèse de recherche et qui seraient susceptibles d'expliquer également le rejet de l'hypothèse nulle. En d'autres termes, rejeter l'hypothèse nulle revient à admettre qu'« il ne se passe pas rien », mais n'indique pas que ce qui se passe est précisément ce que dit l'hypothèse de recherche.

Hypothèse nulle, hypothèse alternative et hypothèse de recherche

Lorsque l'hypothèse de recherche (celle à laquelle le chercheur s'intéresse réellement) conduit à une prédiction précise du paramètre, il est possible de l'identifier à l'hypothèse nulle et non à l'hypothèse alternative. C'est notamment le cas lorsqu'il s'agit de validation de modèles. Une illustration peut ainsi être trouvée dans le test des modèles de causalité, par exemple un modèle causal expliquant la fidélisation des dirigeants par l'attribution de stock-options. Un autre exemple d'hypothèse de recherche correspondant à l'hypothèse nulle est celui de l'hypothèse selon laquelle la survie des organisations est un phénomène aléatoire. Cette dernière hypothèse s'appuie sur les résultats des nombreux travaux empiriques conduits dans le contexte analytique de l'écologie des populations, qui ont montré une absence de différence dans les taux de survie des jeunes et des vieilles organisations ou des petites et des grandes organisations (Schwab et Starbuck, 2009). Mais, quel que soit le choix adopté par le chercheur pour l'hypothèse de recherche (H0 ou H1), les difficultés demeurent : si on identifie l'hypothèse de recherche à l'hypothèse alternative (cas classique), alors, comme déjà évoqué, il suffit de choisir un échantillon suffisamment grand pour être sûr d'obtenir un résultat favorable (significatif) ; si, en revanche, on identifie l'hypothèse de recherche à l'hypothèse nulle, on se trouve confronté au dilemme suivant (Cohen, 1992 ; Poitevineau, 1998 ; Martinez-Pons, 1999) :

- monter un design de recherche très sensible (par exemple en choisissant un grand échantillon) ; c'est se ramener au cas précédent de rejet presque certain de l'hypothèse de recherche, alors même que celle-ci peut constituer une très bonne approximation de la réalité, voire la meilleure disponible ;
- monter un design de recherche très peu sensible (par exemple en choisissant un petit échantillon) ; c'est permettre une corroboration facile et artificielle de l'hypothèse de recherche.

Probabilité de l’hypothèse ou probabilité des données ?

Les TSS fournissent une probabilité des données observées conditionnellement à la véracité de l’hypothèse nulle ou Pr (DonnéesIH0). Pourtant, ce qui intéresse vraiment le chercheur, c’est plutôt la probabilité des hypothèses (H0 et/ou H1) conditionnellement aux données observées Pr (H0IDonnées) ou Pr (H1IDonnées). Cette position peu naturelle et peu intuitive conduit du reste à des erreurs d’interprétation constantes, comme le mentionnent plusieurs auteurs (Carver, 1978 ; Cohen, 1994 ; Gill, 1999). De nombreux utilisateurs interprètent ainsi les résultats observés Pr (DonnéesIH0) comme la probabilité conditionnelle de l’hypothèse nulle, à savoir Pr (H0IDonnées). Pourtant les valeurs de ces deux probabilités peuvent être très différentes l’une de l’autre, par exemple 0,005 et 0,82 dans un cas célèbre décrit par Falk et Greenbaum (1995). Considérons l’exemple d’une recherche visant à tester l’hypothèse selon laquelle la présence d’un service de planification stratégique dans les PME améliore la performance de ces dernières. L’erreur consisterait à confondre les deux probabilités suivantes : 1) Pr (Performance supérieure I Présence d’un service de planification stratégique) qui est la probabilité d’observer une performance supérieure en cas de présence d’un service de planification stratégique dans la PME ; 2) Pr (Présence d’un service de planification stratégique I Performance supérieure) qui est la probabilité de la présence d’un service de planification stratégique dans une PME présentant une performance élevée. Or ces deux probabilités peuvent bien entendu être très différentes.

α est arbitraire, la valeur p est ambiguë

Le seuil de signification α retenu (généralement 5 %, mais parfois aussi 10 % ou 1 %, etc.) est totalement arbitraire et, pourtant, il conduit directement à décider du rejet ou du non-rejet des hypothèses. Prenons l’exemple de la comparaison d’une moyenne d’un échantillon à une valeur donnée. On dispose d’un échantillon constitué de 144 observations. La moyenne trouvée sur cet échantillon est m = 493. L’écart-type estimé sur l’échantillon est s = 46,89 et le seuil de signification observé ou valeur p = 0,07. Dans cet exemple, on ne rejettera pas l’hypothèse (nulle) selon laquelle la moyenne de la population est μ0 = 500 si le seuil de signification α retenu est 5 % alors qu’on la rejettera si le seuil de signification α retenu est 10 %. Ce qui est clairement arbitraire.

La valeur p (ou « seuil de signification observé »), qui est caractéristique des données observées, est indicatrice du degré de réfutation de l’hypothèse nulle : si p est jugée suffisamment faible, on rejette l’hypothèse nulle, on considère qu’on a réussi à en montrer la fausseté et le résultat est déclaré significatif (Poitevineau, 1998 ; Nickerson, 2000). Seulement, la valeur p elle-même n’est pas exempte de critiques. En particulier, elle dépend de la taille de l’échantillon : plus l’échantillon est grand, plus les valeurs p seront faibles, toutes choses égales par ailleurs. Il devient donc difficile de distinguer, dans une valeur p donnée, ce qui provient de la grandeur de l’effet testé (effect size) de ce qui provient de l’effet taille de l’échantillon (sample size). Dans tous les cas, qu’il s’agisse du seuil de signification α ou de la valeur p, la position du curseur qui va déterminer la frontière entre ce qui est – sta-

tistiquement – « significatif » et ce qui ne l'est pas restera une position (valeur) arbitraire.

La grandeur de l'effet

La grandeur de l'effet (effect size) mesure l'amplitude ou la force de la relation entre deux ou plusieurs variables dans la population. Par exemple, une augmentation de 20 % des primes versées aux salariés peut se traduire par une augmentation de 30 % de leur productivité. Les tests ont souvent négligé la grandeur de l'effet (Cohen, 1988, 1994 ; Schwab et Starbuck, 2009). Or, comme le souligne Poitevineau (1998), un résultat statistiquement significatif n'est qu'une indication de l'existence de l'effet supposé alors qu'un résultat non statistiquement significatif n'est qu'un constat d'ignorance. Aller plus loin sur la seule base du test serait assimiler à tort significativité statistique et grandeur de l'effet. En fait, le chercheur est surtout intéressé par la grandeur et la précision des effets. Prenons l'exemple de l'hypothèse de recherche selon laquelle la formation professionnelle accroît la productivité des salariés. Ici, l'enjeu réel n'est pas tant de savoir si la formation a un effet sur la productivité que de savoir si cet effet est suffisamment important pour justifier d'investir dans un plan de formation. L'usage des TSS aurait conduit à formuler l'hypothèse nulle selon laquelle la formation n'a aucun effet et, si cette hypothèse nulle venait à être rejetée, à interpréter le résultat comme une preuve de l'utilité des plans de formation. Naturellement, une telle conclusion n'est pas strictement justifiée par le résultat du TSS. En somme, les questions pertinentes restent les suivantes : est-ce qu'une différence de moyennes est triviale, faible, moyenne ou forte ? Cette différence fait-elle sens, naturellement ? Est-elle suffisamment importante pour être incluse dans un modèle plus large ? Toutes ces questions de recherche importantes ne sont pas traitées par les TSS car elles en dépassent le cadre.

Les chercheurs en sciences sociales ont souvent beaucoup de difficultés à définir ce qui est substantiellement significatif. Autrement dit, ils ont du mal à déterminer à partir de quelle taille une différence, une proportion, un coefficient de corrélation ou de régression, etc. devient important, intéressant, pertinent. Bien qu'il n'existe pas de règles absolues pour interpréter la grandeur des effets, des lignes directrices sont développées dans les différents champs de recherche. À cet égard, l'ouvrage de Cohen (1988) demeure une référence incontournable. L'auteur a défini des conventions qui, au fil des années, sont pratiquement devenues des normes universelles (petit effet = 0,20 ; moyen effet = 0,50 ; grand effet = 0,80). Pourtant, la lecture attentive de l'ouvrage permet de découvrir que l'auteur met en garde contre cette situation et qu'il demande aux chercheurs de définir eux-mêmes la taille de l'effet. « Les valeurs choisies n'ont pas plus de fiabilité comme base que ma propre intuition. Elles sont fournies comme convention parce qu'elles sont nécessaires dans un climat de recherche caractérisé par un manque d'attention pour tout ce qui concerne la grandeur de l'effet. » (Cohen, 1988 : 532)

DES PROBLÈMES D'ORDRE TECHNIQUE : DE FRÉQUENTES ERREURS À L'USAGE

Au-delà de leurs défauts intrinsèques, on reproche aux TSS d'être la source d'erreurs fréquentes commises par les chercheurs, même si on pourrait estimer que ces derniers sont le plus à blâmer. Voici les principales erreurs rencontrées (Sawyer et Peter, 1983 ; Poitevineau, 2004) :

Le renversement des conditions

Une première erreur, la plus typique, et que nous avons déjà évoquée, est de considérer la valeur p (seuil observé) ou α comme une probabilité concernant l'hypothèse (Pr (H0IDonnées)) et non plus conditionnelle à celle-ci Pr (DonnéesIH0). De nombreux auteurs (Cohen, 1994 ; Poitevineau, 1998 ; Gill, 1999) mettent en garde contre les affirmations erronées suivantes :

- « la probabilité que l'hypothèse nulle soit vraie est p (ou α) » ;
- « la probabilité que les résultats soient dus au seul hasard est p (ou α) » ;
- « la probabilité que l'hypothèse alternative soit vraie est 1– p (ou 1– α) ».

La valeur p (seuil observé) est calculée en postulant que l'hypothèse nulle est vraie, c'est-à-dire que toute différence observée est uniquement due au hasard. Cette valeur p sert ensuite à décider du rejet ou du non-rejet de cette hypothèse nulle postulée. Par exemple, un test de différence de scores moyens entre deux échantillons montre une différence de 10 points à laquelle est associée une valeur p de 3 %. Cette valeur p de 3 % signifie que, si l'hypothèse nulle est vraie (c'est-à-dire si les populations ont en fait le même score moyen), il existe 3 chances sur 100 d'observer une différence des scores moyens supérieure ou égale à 10 points et 97 chances sur 100 d'obtenir une différence des scores moyens inférieure ou égale à 10 points entre deux échantillons quelconques tirés de chacune des deux populations. La valeur p porte donc clairement sur la probabilité des données (la différence de scores moyens observée) et non sur la probabilité de l'hypothèse nulle.

Les erreurs d'interprétation s'expliquent par l'écart entre ce que les utilisateurs attendent et ce que les tests fournissent (Cohen, 1994). Les utilisateurs ont recours aux TSS pour décider si les résultats obtenus confirment ou infirment leur hypothèse. Or les tests indiquent la probabilité p d'obtenir les résultats observés (sachant que l'hypothèse nulle est vraie) et non la probabilité de l'hypothèse nulle (au regard des données). Nous avons déjà signalé que les valeurs de ces deux probabilités pouvaient différer très sensiblement même s'il reste vrai que plus p est petit, plus les preuves contre l'hypothèse nulle sont grandes. En termes très généraux, un résultat statistiquement significatif est un résultat qui advient très rarement lorsque l'hypothèse nulle est vraie (Sawyer et Peter, 1983).

On trouve, du reste, des erreurs similaires concernant les intervalles de confiance (ou « fourchettes »). Souvent, les chercheurs affirment à propos d'un intervalle de confiance à 95 % [X ; Y] : « le paramètre π a une probabilité 0,95 de se trouver dans la fourchette (ou l'intervalle)

[X ; Y] ». Cette interprétation naturelle est néanmoins fausse. Les paramètres sont généralement inconnus mais ils ont une valeur fixe, non aléatoire. L'événement « X < π < Y » est vrai ou faux (car π est fixé), et on ne peut pas lui attribuer de probabilité (sinon 1 ou 0). L'interprétation correcte de l'intervalle de confiance 95 % est la suivante : « 95 % des intervalles calculés sur l'ensemble des échantillons possibles (tous ceux qu'il est possible de tirer dans la population) contiennent la vraie valeur π. » Chaque intervalle particulier a une probabilité 0 ou 1 de contenir la vraie valeur. Ici, ce n'est pas le paramètre qui est aléatoire mais les bornes de l'intervalle de confiance (ou « fourchette ») qui varient d'un échantillon à l'autre.

La probabilité de reproduction du résultat

Une deuxième erreur est de considérer 1 – p (ou 1 – α) comme la probabilité de reproduire le résultat observé (Falk et Greenbaum, 1995 ; Poitevineau, 1998 ; Gill, 1999). Les théories traditionnelles des tests ne fournissent aucune indication de la probabilité de reproduire le résultat observé. Cette reproductibilité peut être envisagée de deux manières :

- soit elle ne concerne que la significativité du résultat, c'est le cas le plus courant ; ce qui donne un énoncé du type : « la probabilité qu'une réplication soit significative est 1 – p (ou 1 – α) » ;
- soit la reproductibilité concerne la valeur même de l'effet : « il y a 95 % de chances d'observer un même résultat dans les travaux ultérieurs ».

Ces énoncés sont bien entendus erronés, même s'il est exact que plus le seuil observé p est faible et plus la reproductibilité est assurée. En fait, la reproductibilité des résultats dépend essentiellement de la contrôlabilité des variables pertinentes, comme c'est le cas pour les expérimentations. Par exemple, une recherche peut avoir mis en évidence une relation positive statistiquement significative entre l'adoption d'une technique particulière de formation et la productivité du personnel. Ce résultat peut ne pas se reproduire dans d'autres recherches tout simplement parce qu'une variable aussi importante que la qualité du formateur n'a pas été contrôlée alors qu'elle était la véritable cause de l'amélioration de la productivité, davantage que la technique de formation elle-même. Clairement, rien dans la logique des TSS n'autorise à interpréter un résultat statistiquement significatif comme une indication directe de la probabilité de reproduction du résultat (Carver, 1978). Un autre argument mis en avant par Armstrong (2007b) nous inspire l'illustration suivante : supposons qu'il existe une corrélation de 0,3 entre l'adoption d'un plan de formation et l'accroissement de la productivité du personnel. Si cinquante tentatives de réplication avaient lieu à l'aide de tests ayant une puissance de 50 % (une valeur raisonnable en sciences sociales), alors environ la moitié des tentatives de réplication concluraient à l'existence d'une corrélation et l'autre moitié à l'absence de corrélation, au seuil de 5 %. Dans ce cas, un TSS effectué sur une méta-analyse des 50 réplications conduirait à conclure, à tort, qu'il n'existe pas de corrélation statistiquement significative entre l'adoption d'une technique particulière de formation et la productivité du personnel.

M@n@gement vol. 13 no. 2, 2010, 99-127

Significativité statistique et significativité substantielle

Une troisième erreur est de confondre la significativité statistique avec la significativité substantielle (Sawyer et Peter, 1983 ; Gliner, Morgan, Leech et Harmon, 2001). Cela revient à considérer que plus un résultat est statistiquement significatif, plus il est scientifiquement intéressant, et / ou plus l'effet correspondant dans la population est grand. Les développements précédemment consacrés à la grandeur de l'effet (effect size) et à l'effet de la taille de l'échantillon (sample size) ont montré la différence entre les deux significativités. Rappelons, après Carver (1978), qu'un résultat statistiquement significatif est littéralement un résultat dont la probabilité d'occurrence est faible si l'hypothèse nulle est vraie. Or, comme la significativité statistique dépend de la taille de l'échantillon, des différences triviales sont souvent interprétées comme importantes lorsque la taille de l'échantillon est très grande. Par exemple, supposons que la productivité des salariés soit mesurée sur une échelle allant de 0 à 100 et qu'on dispose de deux échantillons regroupant, l'un, des salariés ayant suivi un plan de formation spécifique et l'autre, des salariés n'ayant pas suivi de plan de formation. Une différence d'un centième de point est constatée entre les scores de productivité moyenne des deux échantillons. Une telle différence d'un centième de point dans un score de performance variant entre 0 et 100 est clairement triviale. Pourtant, cette différence triviale sera statistiquement significative si chaque échantillon est composé de milliers de salariés. À l'inverse, une différence substantielle (par exemple de plusieurs points) peut se révéler statistiquement non significative pour peu que les échantillons soient de taille suffisamment faible (quelques individus). Dans ce dernier cas, Carver (1978) suggère de tenter une réplication pour vérifier si on trouve à nouveau un effet d'intensité comparable. En définitive, il revient au chercheur – qui a réfléchi à son hypothèse de recherche – d'identifier ce qui fait sens. Par exemple, c'est à lui de savoir si, du point de vue de la signification, il ne vaut pas mieux choisir pour hypothèse nulle qu'entre les moyennes de deux groupes la différence est égale à une constante non nulle plutôt qu'à une différence strictement nulle. En tout état de cause, et sur un plan plus général, plusieurs auteurs (Carver, 1993 ; Thompson, 1996) insistent sur l'expression « statistiquement significatif » qui ne devrait jamais être remplacée par l'expression « significatif » tout court.

L'acceptation de l'hypothèse nulle

Une quatrième erreur est de conclure à la véracité de l'hypothèse nulle en cas de résultat non significatif. Par exemple, un chercheur examine l'hypothèse de recherche selon laquelle « la mise en place d'un plan de formation augmente la performance du personnel ». Un test de différence de moyennes le conduit à observer des résultats statistiquement non significatifs : il n'observe pas de différence statistiquement significative entre les groupes ayant suivi ou non un plan de formation. Il doit en conclure qu'il ne peut pas rejeter l'hypothèse nulle et ne doit surtout pas affirmer qu'il « accepte » cette hypothèse nulle. Encore une fois, il peut simplement dire qu'il ne peut pas refuser l'hypothèse nulle. En revanche, il ne doit pas affirmer : « la mise en place d'un plan de formation n'augmente pas la performance du personnel ». Cela équi-

vaut à accepter l'hypothèse nulle et signifie qu'on conclut par inférence (à l'ensemble de la population) à l'absence d'effet, sans se contenter de jugements descriptifs incontestables. De même, la phrase « la valeur 0 étant comprise dans l'intervalle de confiance, on ne peut pas refuser l'hypothèse nulle selon laquelle les deux séries de valeurs ont la même moyenne » est correcte car il s'agit d'une description concernant les échantillons. Cependant, on ne doit pas affirmer : « la distribution de stock-options ne modifie pas la fidélisation des dirigeants » car il s'agit là d'une inférence. De fait, Cohen (1994) a montré que prouver l'hypothèse nulle est une impossibilité logique dans le contexte des TSS. Plusieurs autres auteurs mettent également en garde contre une telle erreur (Gill, 1999 ; Krueger, 2001).

Les erreurs précédentes sonnent comme autant de critiques supplémentaires des TSS : une méthode donnant lieu à tant d'erreurs dans son application, même chez des usagers avertis, n'aurait-elle rien à se reprocher ? Mais les TSS présentent encore un troisième type de problèmes, d'ordre philosophique cette fois.

DES PROBLÈMES D'ORDRE PHILOSOPHIQUE : LES LIMITES DE L'INFÉRENCE

La logique des TSS est fondée sur l'inférence (Fisher, 1942 ; Krueger, 2001 ; Morgan, 2003). Les précurseurs de l'inférence, tels David Hume au milieu du XVIIIe siècle et Karl Pearson au début du XXe siècle, affirment que les futurs événements peuvent être anticipés à partir des séquences ou fréquences d'événements antérieurs (Alexander, 1972 ; MacNabb, 1972 ; Morgan, 2003). D'après Hume et Pearson, les causes et les effets ne peuvent se justifier que comme une série étendue de coïncidences que l'on commence à associer à une anticipation (Black, 1972). L'une des implications philosophiques de cette conception est que, bien qu'il soit possible de prouver que quelque chose est faux (du fait d'une absence de coïncidences), il devient impossible de prouver que quelque chose est vrai (Howell, 1997 ; Krueger, 2001 ; Morgan, 2003).

Fisher, comme Hume et Pearson, estime que l'inférence est fondamentalement le seul processus permettant la découverte de nouvelles connaissances (Fisher, 1942). Par conséquent, l'objectif du système d'inférence de Fisher est de tester – et, plus précisément, de réfuter – une hypothèse selon laquelle un traitement particulier a conduit à des différences entre échantillons (Mulaik et al., 1997 ; Morgan, 2003). Constitutive de l'inférence, la supposition selon laquelle l'avenir ressemble au passé n'est fondée sur aucun argument : elle dérive simplement de l'habitude, par laquelle nous sommes déterminés à attendre, pour l'avenir, le train de choses auquel nous avons été habitués (Hume, 1978). Et comme le montre Morgan (2003), l'analyse de données expérimentales conduit à des inférences sur la probabilité d'événements futurs : lorsque les différences entre les conditions sont improbables sous l'hypothèse nulle, les chercheurs attribuent ces différences à la stabilité des causes sous-jacentes et s'attendent donc à observer de nouveau les mêmes différences dans des circonstances similaires.

La démarche générale consiste à faire une hypothèse de base, à observer un phénomène réel et, ensuite, à évaluer la compatibilité de l'hypothèse de base avec le phénomène réel. Plus précisément, le raisonnement correspond au syllogisme suivant (Gill, 1999) :

1. Si A alors B ;
2. Non B est observé ;
3. Donc non A

Pour les TSS, ce raisonnement se traduit de la manière suivante :

1. Si H0 est vraie alors les données auront une forme particulière ;
2. Les données n'ont pas la forme particulière correspondante ;
3. Donc H0 est fausse.

Le grand problème est que, dans l'application concrète de cette logique formelle aux TSS, on bascule d'affirmations certaines à des affirmations probabilistes. En effet, le raisonnement devient :

1. Si A alors B est hautement probable ;
2. Non B est observé ;
3. Donc A est hautement improbable

Pour les TSS, cela donne :

1. Si H0 est vraie alors les données auront très probablement une forme particulière ;
2. Les données n'ont pas la forme particulière correspondante ;
3. Donc H0 est très probablement fausse.

De prime abord, cette logique semble plausible. Pourtant, il est faux d'affirmer que la présence de données atypiques (ou improbables) sous une hypothèse de base donnée implique que l'hypothèse de base est fausse. Il se peut, tout simplement, qu'un phénomène rare – c'est-à-dire peu probable – soit advenu (Cohen, 1994 ; Gill, 1999). Considérons l'exemple suivant :

1. Si une personne parle français, elle n'est très probablement pas membre de l'AIMS ;
2. La personne est membre de l'AIMS ;
3. Donc elle ne parle très probablement pas français.

Cet exemple, avec sa conclusion absurde, montre bien les limites de l'inférence… statistique fondée sur les TSS. Lorsque l'hypothèse nulle est rejetée dans le cadre d'un TSS, cela suggère simplement que les résultats ne doivent pas être attribués au hasard. Autrement dit, cela suggère qu'« il n'y a pas rien » selon l'expression de Dawes (1991 : 252). Cette inférence probabiliste est une démonstration par l'absurde (modus tollens). Si l'hypothèse nulle est vraie, l'existence de données ordonnées est improbable. Si les données semblent improbables, alors l'hypothèse nulle est probablement fausse. Si l'hypothèse nulle est fausse, alors quelque chose de fondamental – et autre que le hasard – est probablement en œuvre (Chow, 1998 ; Morgan, 2003). Mais le problème principal avec cette chaîne d'inférence, c'est que les syllogismes ne sont pas valides lorsqu'ils sont appliqués à l'inférence. Trois critiques sont principalement avancées (Morgan, 2003). Premièrement, toute hypothèse ponctuelle est fausse, et aucune donnée n'est, à la limite, nécessaire pour la rejeter (Tukey, 1991 ; Morgan, 2003). Par conséquent, l'objectif des TSS doit être autre chose que le

seul rejet d'hypothèses nulles. Deuxièmement, même si l'on suppose qu'une hypothèse est vraie, la probabilité des données compte tenu de cette hypothèse ne correspond pas à la probabilité de cette hypothèse compte tenu des données. Par exemple, la probabilité qu'une entreprise maîtrisant exactement dix facteurs clés de succès soit performante n'est pas nécessairement égale à la probabilité qu'une entreprise performante maîtrise exactement dix facteurs clés de succès. La première probabilité serait en toute hypothèse très forte et la seconde en toute hypothèse très faible. Aucune contradiction, aussi peu probable qu'elle soit, ne peut réfuter quoi que ce soit si les prémisses sont incertaines. Troisièmement, enfin, les TSS ne sont pas d'une aide particulière concernant les possibilités de réplication des résultats dans le futur (Carver, 1978 ; Morgan, 2003).

Hume (1978) a remarqué que l'inférence ne peut pas être autrement validée que par l'inférence elle-même. L'inférence à partir d'un échantillon – de quelque taille qu'il soit – ne peut pas fournir de connaissance certaine à propos des caractéristiques de la population. Pourtant, parce nos inférences ont pu être avérées dans le passé, nous espérons qu'il en sera de même à l'avenir. Cela, en soi, est une inférence qui ne peut être justifiée que par d'autres inférences, et ainsi de suite (Krueger, 2001 ; Morgan, 2003). La recherche empirique doit soit accepter cet acte de foi, soit se briser. Parce que le savoir « doit contenir des prévisions fiables » (Reichenbach, 1951 : 89), nous agissons « comme si nous avions résolu le problème de l'induction » (Dawes, 1997 : 387).

Il est intéressant de constater que c'est une réflexion sur les TSS qui permet de (re)découvrir que l'activité du chercheur – au sein de laquelle la démarche d'inférence occupe une place centrale – exige fondamentalement… des actes de foi (Hume, 1978 ; Reichenbach, 1951 ; Dawes, 1997 ; Krueger, 2001). Car le recours aux TSS et, plus généralement, la préoccupation de généralisation statistique relèvent essentiellement d'une certaine conception – plutôt positiviste – de la science qui reste discutable si l'on envisage la recherche en management comme une « science historique » ayant pour objectif de rendre intelligible une certaine catégorie de phénomènes sans prétendre identifier des « lois » qui les gouverneraient. Dans cette optique, le principal intérêt de la réflexion sur les TSS reste bien de faire apparaître ces derniers comme un outil ordinaire… d'inférence. Tout cela renvoie aux qualités idéales du chercheur (en management) et à sa capacité de trouver un équilibre – toujours fragile, parce que dynamique – entre, d'une part, audace, volonté d'invention, de création, de découverte et, d'autre part, prudence, hésitation, doute, humilité et respect des données de l'expérience sensible qui inspire et éprouve à la fois, en première et dernière instances, toutes les théories et hypothèses du chercheur (en management).

M@n@gement vol. 13 no. 2, 2010, 99-127

COMMENT MIEUX UTILISER LES TESTS DE SIGNIFICATION STATISTIQUE

CERNER LES CAUSES DE LA PERSISTANCE DES PROBLÈMES

En dépit des trois types de problèmes posés par les TSS (statistique, technique, philosophique), leur popularité reste grande auprès des chercheurs (Armstrong, 2007b ; Levine, Weber, Hullett, Hee Sun Park et Lindsey, 2008). Malgré les nombreuses critiques dont leur usage fait constamment l'objet, les TSS sont conventionnellement acceptés comme une preuve de la validité des conclusions et sont une norme incontournable pour la publication des résultats de recherche. Tout se passe comme si on était en présence d'une pratique critiquable aux plans théorique et méthodologique mais sociologiquement adaptée, d'un outil mal utilisé car son mode d'emploi se révèle particulièrement trompeur mais bénéficiant néanmoins d'une aura jusque-là intacte. Poitevineau (1998 ; 2004) résume les principales raisons de ce paradoxe apparent :

- L'ambiguïté de la terminologie : les TSS sont des « tests de signification », ce qui renvoie à « significatif », à quelque chose qui donne du sens, qui a de l'importance, etc. Ce faisant, la confusion entre significativités statistique et substantielle est induite.
- L'objectivité : les chercheurs souhaitent disposer de méthodes objectives et formalisées leur permettant de savoir si un jeu de données présente des variations aléatoires ou systématiques. Et ils estiment important de ne pas devoir s'en remettre à leurs seules intuition et subjectivité pour déterminer la part d'aléatoire et de systématique dans les données. Dès lors, les TSS confèrent aux conclusions des chercheurs cette impression d'objectivité qui est chez eux un souci crucial.
- La scientificité : dans des disciplines comme le management qui souffrent plus ou moins d'un complexe de non-scientificité, du moins par rapport à des sciences plus « dures », l'appareillage mathématique et le formalisme des TSS fournissent à bon compte une apparence de scientificité. En outre, la rigueur des mathématiques et l'aura dont elles jouissent sont censées diffuser sur l'ensemble de la recherche, assurant de facto sa validité.
- Le renfort de Karl Popper : les TSS offrent une grande ressemblance avec l'idée de Popper selon laquelle la démarcation entre énoncés scientifiques et non scientifiques est réalisée sur la base du caractère réfutable, ou non, de ces énoncés. Une hypothèse scientifique est une hypothèse qui peut être empiriquement « testée ». La théorie des TSS a ainsi pu bénéficier du succès des idées de Popper.
- Un confort assuré et une économie : les TSS assurent un confort certain à leurs utilisateurs. Avec leur pouvoir de déclarer « significatif » un effet, les TSS sont vus comme une solution,

déchargeant le chercheur de la tâche d'interprétation, comme si la significativité statistique se suffisait à elle-même.

Tout porte donc à croire que le succès persistant des TSS est dû à un formidable malentendu : une apparence d'objectivité et de scientificité ainsi qu'une illusion d'adéquation aux besoins des chercheurs permise par l'ignorance que la plupart de ces derniers ont de la nature et des conditions d'utilisation desdits TSS. Pourtant, les critiques – qui ne sont pas nouvelles – commencent à produire lentement leurs premiers effets. L'élément déclencheur est non pas un réveil brusque ou une prise de conscience subite des chercheurs utilisateurs des TSS, mais plutôt le relais des critiques pris, ces dernières années, par des institutions comme l'American Psychological Association ou des comités éditoriaux de revues scientifiques qui prescrivent de nouvelles normes de publication. Pour l'essentiel, les résultats des analyses statistiques traditionnelles devraient être complétés – au-delà des seuls seuils de signification observés ou valeurs-p – pour inclure systématiquement la présentation d'indicateurs de la grandeur des effets et leurs estimations par intervalles. Dans cette veine, quelles sont, plus généralement, les principales voies d'amélioration possibles ?

QUELQUES VOIES D'AMÉLIORATION

Plusieurs voies d'amélioration sont envisageables. Une façon pratique de les aborder est de commencer par les recommandations du groupe de travail chargé par le bureau des affaires scientifiques de l'American Psychological Association (APA) d'étudier le rôle des TSS dans la recherche en psychologie (APA, 1996). On pourra ensuite explorer des voies complémentaires.

Les recommandations de l'American Psychological Association

- Tests d'hypothèses : il est difficile d'imaginer une seule situation dans laquelle une décision binaire d'acceptation / refus serait préférable au fait de reporter les valeurs p ou, mieux encore, un intervalle de confiance. Par ailleurs, ne jamais utiliser l'expression malheureuse : « accepter l'hypothèse nulle ».
- Intervalles : des intervalles devraient être fournis pour toute grandeur d'effet concernant les résultats principaux. Fournir de tels intervalles pour les corrélations et les indices d'association ou de variation chaque fois que c'est possible.
- Grandeur des effets : toujours présenter les grandeurs d'effets pour les résultats bruts. Si les unités de mesure ont un sens pratique (par exemple, nombre de cigarettes fumées par jour), préférer une mesure non standardisée (coefficient de régression ou différence de moyennes) à une mesure standardisée.
- Puissance et taille de l'échantillon : fournir l'information sur la taille de l'échantillon et le processus qui a conduit au choix d'une telle taille. Expliciter les postulats concernant la grandeur des effets, l'échantillonnage et la mesure des variables de même que les procédures analytiques utilisées pour le calcul de la puissance. Dans la mesure où le calcul de la puissance fait davantage sens lorsqu'il est effectué avant la collecte et l'examen des don-

 M@n@gement vol. 13 no. 2, 2010, 99-127

nées, il est important de montrer comment des estimations de la grandeur des effets ont été déduites des recherches et théories antérieures pour écarter le soupçon qu'elles ont pu être extraites des données de l'étude en cours ou, pis encore, qu'elles ont été construites pour justifier un échantillon donné.

Les méthodes statistiques complémentaires

De nombreux statisticiens plaident – depuis plus ou moins longtemps – pour des méthodes statistiques alternatives aux TSS « fréquentistes » classiques (Gill, 1999 ; Nickerson, 2000). Parmi ces méthodes statistiques alternatives aux TSS, on peut mentionner les méthodes de vraisemblance et les méthodes bayésiennes (Poitevineau, 1998).

• Les méthodes de vraisemblance : dans le cas simple de deux hypothèses ponctuelles H0 et H1, la méthode du rapport de vraisemblance consiste à calculer le rapport des densités de probabilité de la statistique observée (x) sous H0 et sous H1, à savoir $f(x|H0) / f(x|H1)$. Ce rapport exprime, sur la base des résultats observés, les « chances » d'une hypothèse relativement à l'autre. On peut éventuellement retenir H0 ou H1 selon que ce rapport est supérieur ou inférieur à une constante choisie arbitrairement (un, par exemple, si l'on ne privilégie aucune hypothèse). La méthode du rapport de vraisemblance présente l'avantage de ne faire intervenir ni probabilités a priori, ni éléments non observés. Si le rapport de vraisemblance permet de juger de la force probante des données entre deux hypothèses ponctuelles, il est malheureusement très rare, dans la pratique, que le chercheur soit confronté à un tel cas.

• Les méthodes bayésiennes : utilisée en tant que méthode d'inférence statistique, la méthode bayésienne consiste à calculer, au moyen du théorème de Bayes, la distribution a posteriori pour le paramètre auquel on s'intéresse, à partir :

– des données observées ;
– d'un modèle d'échantillonnage associé ;
– des probabilités a priori sur le paramètre.

Plusieurs auteurs ont préconisé le remplacement des TSS classiques fréquentistes par une approche bayésienne (Edwards, Lindman et Savage, 1963 ; Rouanet, 1996). Au contraire des TSS classiques fréquentistes, l'approche bayésienne concerne directement la probabilité de véracité de l'hypothèse de recherche (Bakan, 1966 ; Carver, 1978). L'approche bayésienne a été, et est encore, beaucoup critiquée comme une méthode trop subjective car elle nécessite de spécifier des probabilités a priori. Toutefois, le poids de la distribution a priori dans la distribution a posteriori diminue d'autant que la masse des données s'accroît. Ainsi, deux chercheurs partant de distributions a priori différentes s'accorderont sur leurs conclusions si les données sont suffisantes. Il est d'ailleurs recommandé de faire varier les distributions a priori (position optimiste, neutre, pessimiste) et d'analyser la sensibilité des résultats. En fait, les méthodes bayésiennes semblent disposer de nombreux atouts pour s'imposer comme véritables « challengers » des TSS. On trouve, du reste, de nombreux exemples d'utilisation des méthodes bayésiennes en sciences de gestion, notamment en fi-

nance (Corless, 1972 ; Holt et Morrow, 1992 ; Sarkar et Sriram, 2001) et en marketing (Roberts, 1963 ; Levitt, 1972). La recherche d'Albert, Grenier, Denis et Rousseau (2008) consacrée à l'étude du risque alimentaire est également totalement pertinente pour les chercheurs en gestion. Par ailleurs, on constate que de plus en plus de logiciels statistiques incorporent désormais des modules d'analyse bayésienne (c'est le cas, par exemple, du logiciel SPSS avec son programme Amos ou encore, plus récemment, du logiciel MPLUS).

Au-delà de la compréhension des recommandations des institutions comme l'American Psychological Association ou de la considération de méthodes nouvelles d'inférence statistique (comme les méthodes bayésiennes), une troisième voie d'amélioration qui nous paraît, de loin, la plus importante, concerne l'attitude même du chercheur.

Reprendre la posture du chercheur

Il s'agit de retrouver certaines qualités fondamentales du chercheur comme l'esprit critique, la vigilance, le doute, l'audace, la créativité, la puissance de la volonté, etc.

Diverses raisons – sociologiques, historiques, cognitives, affectives, etc. – peuvent conduire le chercheur à parfois manquer de prise de distance et d'esprit critique vis-à-vis de son environnement de travail, en particulier vis-à-vis des outils de recherche disponibles. Le type de formation à la recherche reçu (école de pensée, profil des maîtres et des pairs), l'orientation paradigmatique dominante dans les structures d'appartenance (laboratoires, clubs, associations académiques, etc.), de même que les préférences ou aptitudes personnelles, sont clairement structurants et façonnent, pour une grande part, les croyances et comportements du chercheur. Ces éléments peuvent naturellement favoriser le mimétisme et inhiber l'esprit critique en matière de méthodologie de recherche. Or la meilleure recherche, celle susceptible de produire les résultats les plus intéressants, nécessite indubitablement d'aller parfois au-delà du mimétisme basique et de l'usage routinier des dispositifs les plus usités à un moment donné. Concernant plus spécifiquement les TSS, nous avons certes conscience qu'ils sont de nos jours un critère important dans la sélection des articles soumis à publication, dans le sens où un résultat non significatif a généralement encore très peu de chances d'être publié. Le faible nombre de résultats non significatifs publiés peut d'ailleurs aussi bien résulter d'une politique éditoriale délibérée que d'une sélection ou autocensure opérée par les chercheurs eux-mêmes. Dans tous les cas, il s'ensuit un très faible taux de publication de ces résultats non significatifs. Cette situation peut conduire à des conséquences catastrophiques. Réfléchissons un moment au scénario suivant : plusieurs chercheurs testent, indépendamment les uns des autres, une même hypothèse nulle H0 qui est vraie. Environ 5 % d'entre eux trouveraient un résultat significatif (rejet de H0 au seuil de 5 %) et seraient pratiquement les seuls à même de publier, laissant ainsi croire à la réalité du phénomène étudié (rejet de H0). On se retrouverait alors uniquement avec des résultats faux dans la littérature. Et les tentatives de réplications effectuées par des chercheurs peu audacieux ne feraient qu'aggraver la situation : seuls les résultats statistiquement significatifs seraient à nouveau sélectionnés

et publiés. Quelle assurance avons-nous de ne pas être dans un tel scénario lorsque nous procédons à une revue de la littérature ? Pratiquement aucune. C'est là, encore une fois, une illustration du devoir de vigilance, d'audace et d'esprit critique de la part du chercheur. Et ces qualités sont d'autant plus nécessaires que les outils deviennent plus sophistiqués, nombreux et aisément disponibles. Clairement, cette exigence individuelle gagnerait à être accompagnée d'une action collective visant à encourager la publication de résultats statistiquement non significatifs, ce qui permettrait de réduire un grave travers de l'environnement professionnel des chercheurs. Dans cette perspective, il est intéressant de noter que, dans des disciplines scientifiques autres que le management, certaines revues scientifiques encouragent déjà la publication de résultats non scientifiques : par exemple, les variantes de « Journal of Negative Results ». D'autres revues scientifiques requièrent qu'une expérience soit enregistrée avant d'être entreprise afin d'éviter l'autocensure : par exemple, plusieurs revues membres de l'International Committee of Medical Journal Editors.

CONCLUSION

Cet article a voulu fortement attirer l'attention des chercheurs en management sur les dangers liés à l'usage irréfléchi des TSS. Il prend appui sur une série de publications qui ont nourri et continuent d'alimenter la critique des TSS. Ces publications ont concerné pratiquement tous les champs disciplinaires : la statistique (Berkson, 1942), la psychologie (Hunter, 1997), la sociologie (Selvin, 1957), le marketing (Sawyer et Peter, 1983), la comptabilité (Lindsay, 1995), les sciences politiques (Gill, 1999), les sciences de l'éducation (Morgan, 2003), la psychiatrie (Gliner et al., 2001), la prospective (Armstrong, 2007a), l'écologie (Anderson, Burnham et Thompson, 2000 ; Gibbons, Crout et Healey, 2007), la météorologie (Nicholls, 2001), la communication (Levine et al., 2008), etc.

On remarque que la critique des TSS a connu un nouvel essor à partir de la deuxième moitié des années 1990 et qu'elle s'est progressivement étendue de la statistique à la psychologie avant de traverser pratiquement tous les champs disciplinaires, à l'exception notable du… management qui commence tout juste à soulever la question (Mbengue, 2007 ; Schwab et Starbuck, 2009). C'est précisément de là qu'est née la motivation principale de cet article : informer la communauté des chercheurs en management de l'existence de cette critique des TSS, en détailler le contenu et les enjeux (les dangers liés à l'usage irréfléchi des TSS) et fournir des recommandations permettant d'améliorer la démarche de test d'hypothèses et, plus généralement, la démarche d'inférence – statistique – dans la recherche (en management). Par rapport à l'article de Schwab et Starbuck (2009), notre texte mobilise une littérature plus large, offre une discussion approfondie de l'inférence, réorganise les problèmes posés par les TSS en trois types (statistique, technique et philosophique), fournit plusieurs exemples liés au management

et propose au chercheur de nombreuses recommandations concrètes organisées en trois catégories (le suivi des recommandations de l'APA, l'usage de méthodes statistiques complémentaires ou alternatives aux TSS et le retour aux qualités fondamentales du chercheur).

Il existe un accord général sur les dangers (de l'usage) des TSS. Un premier danger pour le chercheur utilisant les TSS serait d'ignorer leur mode d'emploi, c'est-à-dire leurs conditions d'utilisation. Ce danger devient particulièrement menaçant compte tenu de la disponibilité croissante des logiciels statistiques. Un autre danger pour le chercheur consisterait à s'abriter derrière l'image scientifique des tests statistiques, à céder à leur aura et au confort apparent lié à leur utilisation pour abdiquer sa responsabilité. Or c'est le chercheur qui doit choisir s'il teste ou pas, ce qu'il teste et par quel moyen il le teste. Mais, plus encore, le chercheur doit garder à l'esprit que les TSS ne sont qu'un instrument à l'intérieur d'un dispositif et d'une démarche de recherche : cette recherche commence avant l'éventuel test, se poursuit pendant le test et continue après le test. Quant au test lui-même, il n'est qu'un outil et, en tant que tel, il ne vaut que si on sait s'en servir et à bon escient. De ce point de vue, les questions récurrentes sur l'utilité des TSS sont un bon stimulant et un garde-fou précieux pour l'exercice d'une saine activité de recherche.

Nous avons commencé cet article en rappelant que les TSS étaient au cœur de la statistique inférentielle et, par suite, de la démarche d'inférence (Krueger, 2001 ; Morgan, 2003). Nous avons également montré que jamais méthode statistique n'avait été autant critiquée tout en restant éminemment populaire et très largement utilisée à mauvais escient (Krueger, 2001 ; Armstrong, 2007a, 2007b ; Levine et al., 2008). La question qui était posée dans cet article était de savoir si, en définitive, il fallait, oui ou non, brûler les TSS.

L'analyse de la critique des TSS et la prise de conscience de possibilités alternatives pourraient conduire à répondre par l'affirmative : après tout, l'abolition des TSS ne mettrait apparemment pas en péril la démarche d'inférence, encore moins l'activité de recherche scientifique. Pourtant, il est apparu qu'une grande part des défauts des TSS est liée à l'usage inapproprié qui en est fait, ce qui remet moins en cause l'outil TSS que ses utilisateurs. En ce sens, la mise à mort des TSS s'apparenterait à une sentence pour le moins excessive. Maintenant, on peut évidemment reprocher à un outil de ne pas être suffisamment facile d'utilisation, ce qui pourrait finalement conduire à un verdict intermédiaire entre la peine capitale et l'acquittement.

En fait, la question de l'opportunité de l'incinération des TSS est en soi plus intéressante que n'importe quelle réponse – positive, négative, intermédiaire – qui pourrait lui être apportée. En effet, une telle question renvoie, fondamentalement, à la place de l'inférence dans la démarche de recherche. Seulement, les TSS n'ont pas le monopole de l'inférence statistique, encore moins de la démarche d'inférence en général. C'est pourquoi leur critique pouvait difficilement être analysée en dehors d'une réflexion plus générale sur la nature de l'inférence et son statut dans l'activité des chercheurs, ce qui a été esquissé dans le cadre de cet article.

Finalement, notre texte se révèle moins iconoclaste que son titre ne le laisse présager. Il peut être lu comme une réponse à la question de savoir comment mieux utiliser les TSS. À l'analyse, il n'apparaît pas véritablement nécessaire de proscrire l'usage de ces tests. Tout juste faut-il mettre en garde contre une façon irréfléchie et routinière d'y recourir. Alors que l'utilisation de cet outil a déjà fait l'objet d'une abondante réflexion critique dans la plupart des disciplines scientifiques, ce n'est malheureusement pas le cas dans le domaine du management. L'essentiel du texte vise donc à alerter les chercheurs en management sur les trois types de problèmes – statistique, technique, philosophique – posés par ces TSS. L'article en appelle donc à une utilisation plus rigoureuse, plus consciente, plus réfléchie et plus critique de ces tests, tout en suggérant de recourir, à l'occasion, à d'autres méthodes statistiques, en l'occurrence les méthodes bayésiennes ainsi que les méthodes d'estimation, ponctuelle ou par intervalle.
Bien entendu, notre recherche n'a pas épuisé toutes les questions posées par l'usage des TSS. Plusieurs pistes de recherche future sont ouvertes. Une première piste consisterait à mener une grande enquête quantitative sur les pratiques des chercheurs en management. Certes, on peut trouver raisonnable l'hypothèse selon laquelle le champ du management aurait peu de raisons de différer de l'ensemble des autres champs disciplinaires au sein desquels les enquêtes effectuées ont jusqu'à ce jour produit des résultats constants quant à la prévalence généralisée d'erreurs dans l'usage des TSS. Certes, également, de nombreux éléments qualitatifs nous suggèrent que peu de chercheurs en management ont réellement connaissance de l'existence –et encore moins de la teneur – de la critique des TSS. Cependant, seule une enquête quantitative permettrait de connaître l'étendue précise et la nature exacte du mal ou du risque : par exemple, quelles sont les erreurs les plus fréquentes dans la communauté des chercheurs en management et dans quelles circonstances y est-on le plus exposé ? Une telle enquête pourrait porter sur les articles publiés – ce qui a été la principale démarche adoptée jusqu'à ce jour dans les travaux conduits dans les autres champs disciplinaires – mais également sur les pratiques ou connaissances des chercheurs mesurées à travers des entretiens ou par questionnaires. En ce qui concerne cette seconde méthode, la seule exception, à notre connaissance, est l'étude de Mittag et Thompson (2000). Cette seconde méthode d'enquête nous semble très importante au regard du rôle croissant de la littérature grise – avec le développement d'Internet – et des conférences académiques – avec ou sans publications d'actes – dans la dissémination des bonnes – ou mauvaises – pratiques de recherche. Une deuxième piste de recherche pourrait consister en la conduite de méta-analyses permettant des comparaisons diachroniques et / ou transversales, c'est-à-dire entre champs disciplinaires sur un mode proche du travail effectué – mais cela date de plusieurs décennies déjà – par Morrison et Henkel (1970). Toutes ces grandes enquêtes quantitatives pourraient être utilement combinées avec des études qualitatives fines dans le but d'aboutir à des diagnostics précis pouvant conduire à des pistes

de thérapies efficaces. Nous espérons que de nombreux chercheurs exploreront ces pistes importantes. Mais, surtout, nous espérons qu'ils le feront en mobilisant leurs qualités fondamentales de chercheurs : l'esprit critique, le refus du mimétisme, le culte du doute, la créativité et… la persévérance !

Ababacar MBENGUE est professeur à l'université de Reims et à Reims Management School. Ses thèmes de recherche sont le management stratégique et le renforcement des capacités des organisations, la gestion des connaissances et la méthodologie. Il a été professeur visitant à Wharton (Snider Enterpreneurial Center) et à l'Université d'Orel (Russie).

RÉFÉRENCES

- Albert, I., Grenier, E., Denis, J. B., & Rousseau, J. (2008). Quantitative Risk Assessment from Farm to Fork and Beyond: A Global Bayesian Approach Concerning Food-Borne Diseases. *Risk Analysis, 28*(2), 557-571.
- Alexander, P. (1972). Karl Pearson. The encyclopedia of philosophy. *New York: Macmillian, 6,* 68-69.
- Anderson, D. R, Burnham, K. P., & Thompson, W. L. (2000). Null Hypothesis Testing: Problems, Prevalence, and an Alternative. *Journal of Wildlife Management, 64*(4), 912-923.
- APA (1996). Task Force on Statistical Inference Report. Washington, DC: American Psychological Association.
- Armstrong, J. S. (2007a). Significance tests harm progress in forecasting. *International Journal of Forecasting, 23*(2), 321-327.
- Armstrong, J. S. (2007b). Statistical significance tests are unnecessary even when properly done and properly interpreted: Reply to commentaries. *International Journal of Forecasting, 23*(2), 335-336.
- Baillargeon, G., & Rainville, J. (1978). *Statistique appliquée* (Tome 2, 6e édition). Trois-Rivières: Les Éditions SMG.
- Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66,* 423-437.
- Berkson, J. (1942). Tests of significance considered as evidence. *Journal of the American Statistical Association, 37*, 325-335.
- Black, M. (1972). Induction. The encyclopedia of philosophy. *New York: MacMillan, 4*, 169-181.
- Carver, R. (1978). The case against statistical significance testing. *Harvard Educational Review, 48*, 378-399.
- Carver, R. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education, 61*, 287-292.
- Chow, S. L. (1998). Statistical significance: Rationale, validity and utility. *Behavioral and Brain Sciences, 21*, 169-240.
- Cohen, J. (1988). *Statistical power analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
- Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 55-159.
- Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*, 997-1003.
- Corless, J. C. (1972). Assessing prior distributions for applying Bayesian statistics in auditing. *Accounting Review, 47*, 556-566.
- Cox, D. R. (1958). Some problems connected with statistical inference. *Annals of Mathematical Statistics, 29*, 357-372.
- Dawes, R. M. (1991). Probabilistic versus causal thinking. In D. Cicchetti & W. M. Grove (Eds.), Thinking clearly about psychology: Vol. 1. Matters of public interest: Essays in honor of Paul Everett Meehl (pp. 235-264). Minneapolis: University of Minnesota Press.
- Dawes, R. M. (1997). *Qualitative consistency masquerading as quantitative fit.* In M. L. Dalla Chiara, D. Kees, D. Mundici & J. van Bentheim (Eds.), Structures and norms in science (pp. 387-394). Dordrecht, the Netherlands: Kluwer Academic.
- Dodge, Y. (1993). *Statistique : Dictionnaire encyclopédique.* Paris : Dunod.
- Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70,* 193-242.
- Falk, R., & Greenbaum, C. W. (1995). Significance Tests Die Hard. *Theory and Psychology, 5,* 396-400.
- Firestone, W. A. (1993). Alternative arguments for generalizing from data as applied to qualitative research. *Educational Researcher, 22*(4), 16-23.

- Finch, S., Cumming, G., & Thomason, N. (2001).
Reporting of statistical inference in the Journal of Applied Psychology: Little evidence of reform. *Educational and Psychological Measurement, 61,* 181-210.

- Fisher, R. A. (1942).
The design of experiments (3rd ed.). London: Oliver & Boyd.

- Gibbons, J. M., Crout, N. M., & Healey, J. R. (2007).
What role should null-hypothesis significance tests have in statistical education and hypothesis falsification? *Trends in Ecology & Evolution, 22*(9), 445-446.

- Gill, J. (1999).
The Insignificance of Null Hypothesis Significance Testing. *Political Research Quarterly, 52(*3), 647-674.

- Gliner, J. A., Morgan, G. A., Leech, N. L., & Harmon, R. J. (2001).
Problems with null hypothesis significance tests. Journal of the American *Academy of Child and Adolescent Psychiatry, 40,* 250-252.

- Hagen, R. L. (1997).
In praise of the null hypothesis statistical test. *American Psychologist, 52*, 15-24.

- Harlow, L. L. (1997).
Significance testing introduction and overview. In L. L. Harlow, S. A. Mulaik & J. H. Steiger (Eds.), What if there were no significance tests? (pp. 1-21). Mahwah, NJ: Lawrence Erlbaum.

- Holt, D. L., & Morrow, P.C. (1992).
Risk assessment judgments of auditors and bank lenders: A comparative analysis of conformance to Bayes' theorem. Accounting, *Organizations and Society, 17*(6), 549-559.

- Howell, D. C. (1997).
Statistical methods for psychology (4th ed.). Belmont, CA: Duxbury.

- Hubbard, R., & Bayarri, M. J. (2003).
Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing. *The American Statistician, 57*, 171-178.

- Hume, D. (1978).
A treatise of human nature. Glasgow, Scotland: William Collins (original datant de 1739).

- Hunter, J. (1997).
Needed: A ban on the significance test. *Psychological Science, 8,* 3-7.

- Kanji, G. K. (1993).
100 Structural tests. Thousand Oaks: Sage.

- Kennedy, M. M. (1979).
Generalizing from single case studies. *Evaluation Quarterly, 3*(4), 661-678.

- Krueger, J. (2001).
Null Hypothesis Significance Testing. *American Psychologist, 56*(1), 16-26.

- Levine, T., Weber, R., Hullett, C., Park, H. S., & Lindsey, L. (2008).
A Critical Assessment of Null Hypothesis Significance Testing in Quantitative Communication Research. *Human Communication Research, 34*(2), 171-187.

- Levitt, T. (1972).
Industrial Purchasing Behavior: A Bayesian Reanalysis. *Journal of Business Administration, 4,* 79-81.

- Lindsay, R. M. (1995).
Reconsidering the Status of Tests of Significance: An Alternative Criterion of Adequacy. *Accounting, Organizations and Society, 20*, 35-53.

- MacNabb, D. (1972).
David Hume. *The encyclopedia of philosophy.* New York: MacMillan, 4, 74-90.

- Martinez-Pons, M. (1999).
Statistics in modern research: Applications in the social sciences and education. New York: New York University.

- Mbengue, A. 2007.
Tests statistiques de signification. In R.-A. Thietart & coll. (Eds.), Méthodes de recherche en management (pp. 297-349). Paris: Dunod.

- Meehl, P. E. (1990).
Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 56* (Suppl. 1), 195-244.

- Mittag, K. C., & Thompson, B. (2000).
A national survey of AERA members' perceptions of statistical significance tests and other statistical issues. *Educational Researcher, 29*(4), 14-20.

- Morgan, P. L. (2003).
Null Hypothesis Significance Testing: Philosophical and Practical Considerations of a Statistical Controversy. *Exceptionality, 11*(4), 209-221.

- Morrison, D. E, & Henkel, R. E. (Eds.) (1970).
The Significance Test Controversy. Chicago: Aldine.

- Mulaik, S.A., Raju, N.S., & Harshman, R.A. (1997).
There is a time and place for significance testing. In L. L. Harlow, S. A. Mulaik & J. H. Steiger (Eds.), What if there were no significance tests? (pp. 65-116). Mahwah, NJ: Lawrence Erlbaum.

- Nelson, N., Rosenthal, R., & Rosnow, R. L. (1986).
Interpretation of significance levels and effect sizes by psychological researchers. *American Psychologist, 41,* 1299-1301.

- Nicholls, N. (2001).
The insignificance of significance testing. *Bulletin of the American Meteorological Society, 82,* 981-986.

- Nickerson, R. S. (2000).
Null Hypothesis Significance Testing: A Review of an Old and Continuing Controversy. *Psychological Methods, 5*(2), 241-301.

- Nunnally, J. (1960).
The place of statistics in psychology. *Educational and Psychological Measurement, 20,* 641-650.

- Oakes, M. (1986).
Statistical inference: A commentary for the social and behavioral sciences. New York: Wiley.

- Poitevineau, J. (1998).
Méthodologie de l'analyse des données expérimentales : étude de la pratique des tests statistiques chez les chercheurs en psychologie, approches normative, prescriptive et descriptive. Thèse de Doctorat, Université de Rouen.

- Poitevineau, J. (2004).
L'usage des tests statistiques par les chercheurs en psychologie : aspects normatif, descriptif et prescriptif. *Mathématiques et Sciences Humaines, 167*, 5-25.

- Reichenbach, H. (1951).
The rise of scientific philosophy. Berkeley: University of California Press.

- Roberts, H. V. (1963).
Bayesian Statistics in Marketing. *Journal of Marketing, 27,* 1-4.

- Rozeboom, W. W. (1960).
The Fallacy of The Null-Hypothesis Significance Test. *Psychological Bulletin, 57*(5), 416-428.

- Rouanet, H. (1996).
Bayesian methods for assessing importance of effects. *Psychological Bulletin, 119*, 149-158.

- Sarkar, S., & Sriram, R. S. (2001).
Bayesian Models for Early Warning of Bank Failures. *Management Science, 47*(11), 1457-1475.

- Sawyer, A. G., & Peter, J. P. (1983).
The Significance of Statistical Significance Tests in Marketing Research. *Journal of Marketing Research, 20*(2), 122-133.

- Schmidt, F. L., & Hunter, J. E. (2002).
Are there benefits from NHST? *American Psychologist, 57*, 65-66.

- Schwab, A., & Starbuck, W. H. (2009).
Null-hypothesis significance tests in behavioral and management research: we can do better. *Research Methodology in Strategy and Management, 5*, 29-54.

- Selvin, H. C. (1957).
A Critique of Tests of Significance in Survey Research. *American Sociological Review, 22*, 519-527.

- Sincich, T. (1996).
Business statistics by example. Upper Saddle River, NJ: Prentice-Hall.

- Smaling, A. (2003).
Inductive, analogical, and communicative generalization. *International Journal of Qualitative Methods, 2*(1), 1-31.

- Thompson, B. (1996).
AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25*(2), 26-30.

- Tukey, J. W. (1991).
The philosophy of multiple comparisons. *Statistical Science, 6*, 100-116.

- Wainer, H. (1999).
One cheer for null hypothesis significance testing. *Psychological Methods, 6*, 212-213.

- Yin, R. (1984).
Case study research: Design and methods. Beverly Hills, CA: Sage.

■ Zikmund, W. G. (1994).
Business research methods. Orlando, Florida: The Dryden Press.

■ Zuckerman, M., Hodgins, H. S., Zuckerman, A., & Rosenthal, R. (1993).
Contemporary issues in the analysis of data: A survey of 551 psychologists. *Psychological Science, 4,* 49-53.